VIII. - Les méthodes de corrélation

Les pages qui suivent sont tirées d'une étude inédite que d'ALVERNY *nous avait adressée peu avant sa mort, suivenue en 1930. Nous avons dû la résumer pour la faire tenir dans le cadre que nous nous sommes imposé.*

L. S.

La méthode des corrélations a pour but de préciser la relation entre deux séries de phénomènes mesurés, de rechercher la probabilité d'une cause commune ou d'une relation directe de cause à effet entre eux.

Comme telle, elle est d'un emploi très important pour les sciences naturelles ; mais comme toutes les méthodes mathématiques, elle veut être employée à bon escient. Et il n'est pas inutile de la discuter, en montrant ses conditions logiques et les limites hors desquelles elle travaille à faux.

Nous le ferons sur un exemple concret, sur le problème forestier suivant :

Y a-t-il relation entre la quantité de pluie tombée en été et l'accroissement du sapin ? Ce sont des séries de phénomènes mesurables et que nous avons mesurés en effet de notre mieux, pour plusieurs forêts d'Auvergne.

Le tableau A ci-dessous montre les chiffres des onze années, 1915 à 1925, qui vont servir de matière à notre exposé.

Pourquoi onze ans ? A cause d'une oscillation très apparente pour laquelle, comme en toute matière touchant la météorologie, il faut penser à la période des taches du soleil qui est d'environ 11 ans, si l'on veut avoir une moyenne à peu près juste.

L'accroissement mesuré est la *veine* du bois (accroissement du rayon de l'arbre à 1,30 m en mm), sans tenir compte de l'écorce : il est donné par la moyenne de 146 sondages à la tarière dans des peuplements de sapin qui n'ont pas subi de coupes dans cet intervalle.

La pluie en mm est celle des cinq mois de mai à septembre, donnée par la moyenne de trois stations météorologiques qui encadrent, topographiquement et en altitude, les forêts étudiées.

Pourquoi n'avoir pas pris les faits météorologiques de l'année entière ? Parce qu'on ne peut guère attribuer d'importance à ce qui s'est passé en dehors de la saison de végétation ; et qu'un

premier aperçu a montré, en effet, l'absence de toute corrélation avec les faits moyens de l'année. Mais pourquoi cinq mois ? On aurait peut-être aussi bien fait de n'en prendre que quatre et d'exclure septembre, la saison d'activité du sapin étant le plus souvent ici close à la fin d'août.

Tableau A

Données sur l'épaisseur de la veine annuelle du bois de sapin et sur la hauteur des pluies tombées en été.

Années	Veine	Pluie
1915	3,0 mm	610 mm
1916	2,9	561
1917	2,6	596
1918	2,4	452
1919	2,1	292
1920	2,3	504
1921	1,9	470
1922	1,5	436
1923	1,7	280
1924	1,4	480
1925	2,1	498
Sommes :	23,9	5 179
Moyenne 1/11	2,2 mm	471 mm

Indice de dépendance

Dans l'exemple du tableau A, considérons, pour les comparer, les deux séries : P, pluies d'été et V, veines du bois.

La première idée qui vient à l'esprit est de chercher, pour chacune,

le *sens de la variation* d'une année à l'autre. Désignons par + une augmentation, par — une diminution. Il y aura dépendance probable, si la première série (P) variant dans un sens, la seconde série (V) varie dans le même sens: et l'on jugera la prédominance des réponses favorables ou non à la majorité relative.

Suivant la convention algébrique, si l'on fait pour chaque année

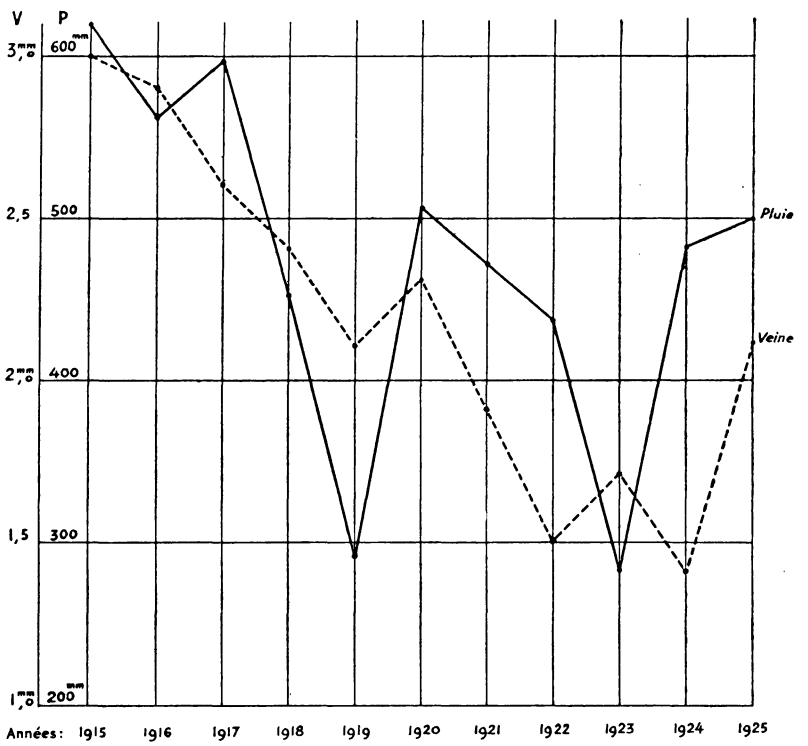

FIG. 16.

Relations entre la hauteur des pluies tombées en été et l'épaisseur de la *veine* annuelle du bois de sapin.

la combinaison des signes de variation, il y a *indice de dépendance directe* à chaque année où le produit est positif (+ + ou — —), indice de dépendance *inverse* (l'accroissement diminuerait lorsqu'il pleut davantage) à chaque année où le produit est négatif (+ — ou — +). Nous obtenons ainsi le tableau B.

TABLEAU B

Indice de dépendance

Années	V	Variation	P	Variation	Produits d	Produits i
	(mm)		(mm)			
1915	3,0		610			
		—		—	+	
1916	2,9		561			
		—		+		—
1917	2,6		596			
		—		—	+	
1918	2,4		452			
		—		—	+	
1919	2,1		292			
		+		+	+	
1920	2,3		504			
		—		—	+	
1921	1,9		470			
		—		—	+	
1922	1,5		436			
		+		—		—
1923	1,7		280			
		—		+		—
1924	1,4		480			
		+		+	+	
1925	2,1		498			
					d = +7	i = —3

Ici, on trouve que 5 fois, il y a eu simultanément diminution et 2 fois simultanément augmentation, au total donc 7 indications directes (d). Les 3 autres variations ont été inverses (i). L'indice de dépendance est d'une façon générale:

$$\frac{d - i}{d + i}$$

Il peut varier de — 1 à + 1. S'il est presque nul, la probabilité de cause à effet est négligeable, invraisemblable. S'il dépasse 0,5, on peut parier pour la dépendance ; on en est sûr, s'il atteint 1.

S'il y a prédominance des dépendances inverses, l'indice est négatif et donne une probabilité de dépendance inverse.

Ici, avec 7 indications directes et 3 inverses, l'indice est :

$$\frac{7 - 3}{7 + 3} = 0{,}4$$

Sans paraître bien fort, il encourage à chercher.

C'est exactement comme si nous avions tracé sur du papier quadrillé les deux lignes brisées Pluie et Veine, année par année, pour comparer le mouvement ascendant ou descendant des parties correspondantes de ces lignes. (fig. 16). L'indication est assez superficielle, car notre indice ne tient pas compte de la grandeur des variations ; un petit mouvement compte autant qu'un grand, et cela n'est pas juste.

Coefficient de corrélation

Considérons toujours les mesures qui se correspondent d'une série à l'autre par paires indépendantes. Mais travaillons cette fois sur les écarts de chaque mesure avec la moyenne de sa série.

Par exemple en 1915, la veine du bois constatée est de 3,00 mm, son écart est + 0,8 au-dessus de la moyenne arithmétique 2,2 mm des mesures prises pour chacune des 11 années consécutives. Et en 1924, la veine la plus faible 1,4 mm a un écart négatif — 0,8 sur cette même moyenne. Pareillement, la moyenne arithmétique des pluies étant 471 mm, l'écart de 1915 est + 139, celui de 1923 est — 191.

Pour avoir des chiffres plus maniables, rien ne s'oppose à un changement d'unité et on peut chiffrer les veines du bois en dixièmes de millimètres et les chutes de pluies en centimètres. Cela équivaut à un changement d'échelle dans les ordonnées des tracés graphiques. C'est ainsi que nous avons exprimé les valeurs du tableau C ci-dessous.

On procède successivement pour chaque paire de mesures (chaque année d'observations) à la multiplication algébrique, en tenant compte des signes, des écarts de la 1re série p par ceux de la seconde v. On range dans une colonne les produits positifs, dans une autre colonne les produits négatifs et l'on fait la somme algébrique. Cette somme montre ici une prépondérance des produits positifs : il y a donc corrélation positive.

Pour comparer cette somme algébrique des produits au produit des sommes des carrés des écarts, on les divise l'une par l'autre. Le coefficient de corrélation *r* est le rapport ainsi obtenu. Le détail du calcul est indiqué sous le tableau C.

TABLEAU C

Coefficient de corrélation de Pearson

Années	Pluie	Ecarts p	p^2	Veine	Ecarts v	v^2	Produits p × v	
							+	—
	cm	cm		1/10 mm	1/10 mm			
1915	61,0	+ 13,9	196	30	+ 8	64	+ 111	
1916	56,1	+ 9,0	81	29	+ 7	49	+ 63	
1917	59,6	+ 12,5	156	26	+ 4	16	+ 50	
1918	45,2	— 1,9	4	24	+ 2	4		— 4
1919	29,2	— 17,9	320	21	— 1	1	+ 18	
1920	50,4	+ 3,3	11	23	+ 1	1	+ 3	
1921	47,0	— 0,1	ε	19	— 3	9	+ ε	
1922	43,6	— 3,5	12	15	— 7	49	+ 25	
1923	28,0	— 19,1	364	17	— 5	25	+ 95	
1924	48,0	+ 0,9	1	14	— 8	64		— 7
1925	49,8	+ 2,7	7	21	— 1	1		— 3
Sommes :			1 152			283	+ 365	— 14
Moyennes :	47,1			22				

$$r = \frac{\Sigma pv}{\sqrt{\Sigma p^2 \times \Sigma v^2}}$$

$$\Sigma pv = + 365 - 14 = + 351$$

$$\sqrt{\Sigma p^2 \times \Sigma v^2} = \sqrt{1152 \times 283} = 571$$

$$r = \frac{+ 351}{571} = + 0{,}61$$

Si la corrélation était absolue, mécaniquement proportionnelle et directe, si par exemple 200 mm de pluie donnaient toujours 1 mm d'accroissement, le coefficient r serait + 1 ; si cette corrélation était absolue, mais inverse, le coefficient r serait — 1.

Entre la certitude (+ 1) d'une corrélation directe et la certitude (— 1) d'une corrélation inverse, le coefficient r peut prendre toutes les valeurs : près de 0, il signifiera l'indifférence, l'absence de relation entre les deux séries de phénomènes, du moins tels qu'ils sont chiffrés ; quand il dépassera 0,5, il y aura une probabilité de plus en plus sérieuse pour une relation de cause à effet ou de cause commune.

Ici, nous obtenons + 0,61, et nous pouvons dire que la relation directe entre la quantité de pluie tombée en été et l'accroissement du sapin est nettement probable.

L. SCHAEFFER.

(d'après A. d'ALVERNY.)